# Modélisation bayésienne hiérarchique pour l'écologie et la recherche environnementale


Frédéric Mortier
**B& SEF**

Pierrette Chagneau
**IRMAR-INSA**

Marie-Pierre Etienne
**AgroParisTech**

Nicolas Picard
**B& SEF**

Cyril Piou
**Bio-agresseurs**

Vivien Rossi
**UMR ECOFOG**



SFDS : Tunis, Mai-2011

# Plan

# Contexte I

1. Face aux enjeux actuels
   - Conservation et gestion de la bio-diversité
   - Les changements climatiques
   - ...
2. Compréhension intégrée
   - Sources d'informations multiples
   - Échelles de temps et d'espace variables
   - Nombreuses sources d'incertitudes

# Contexte II

L'approche hiérarchique bayesienne permet en décomposant un problème en une série de modèles conditionnels plus simples,

1. de prendre en compte différentes sources d'incertitudes
2. d'utiliser des données de différentes résolutions
3. d'intégrer des dépendances spatiales, temporelles ou spatio-temporelles compliquées
4. d'intégrer la connaissance d'experts

# État de l'art

## Articles de synthèse

1. 2003, "Hierarchical Models in Environmental Science", *International statistical review*, [Wik03]
2. 2004, "Bayesian Inference in ecology", *ecology letters*, [Ell04]
3. 2005, "Why environmental scientists are becoming bayesians, *ecology letters* [Cla05]
4. 2006, "A future for models and data in environmental science", *trend in ecology and evolution*, [CG06]
5. 2007, "Hierarchical linear models and the measurement of ecological system", *ecology letters*, [MD07]
6. 2009, "Accounting for uncertainty in ecological analysis", [CCC+09]

# État de l'art

## Articles de synthèse

1. 2003, "Hierarchical Models in Environmental Science", *International statistical review*, [Wik03]
2. 2004, "Bayesian Inference in ecology", *ecology letters*, [Ell04]
3. 2005, "Why environmental scientists are becoming bayesians, *ecology letters* [Cla05]
4. 2006, "A future for models and data in environmental science", *trend in ecology and evolution*, [CG06]
5. 2007, "Hierarchical linear models and the measurement of ecological system", *ecology letters*, [MD07]
6. 2009, "Accounting for uncertainty in ecological analysis", [CCC+09]

## Applications

- Génétique
- Écologie
- Dynamique des populations
- climatologie, paleo-climatologie

# Le principe général

La modélisation hiérarchique se fonde sur un raisonnement conditionnel [WR04, PB07] :
La distribution jointe de 3 variables aléatoires X, Y et Z peut se décomposer comme

$$[X, Y, Z] = [X|Y, Z][Y|Z][Z]$$

Remarque :

- Le choix de cette décomposition est basé sur notre connaissance et sur des hypothèses de simplification du problème
- On modifie les hypothèses d'indépendance par de l'indépendance conditionnelle.
- Il est peut-être plus simple de penser un problème de manière conditionnelle que globalement

# Le principe général et représentation graphique

1. Modèle des données :

$$[obs|\vartheta, \theta_1]$$

2. Modèle des processus :

$$[\vartheta|\theta_2]$$

3. Modèle des paramètres :

$$[\theta_1, \theta_2]$$

# Le principe général et représentation graphique

1. Modèle des données :

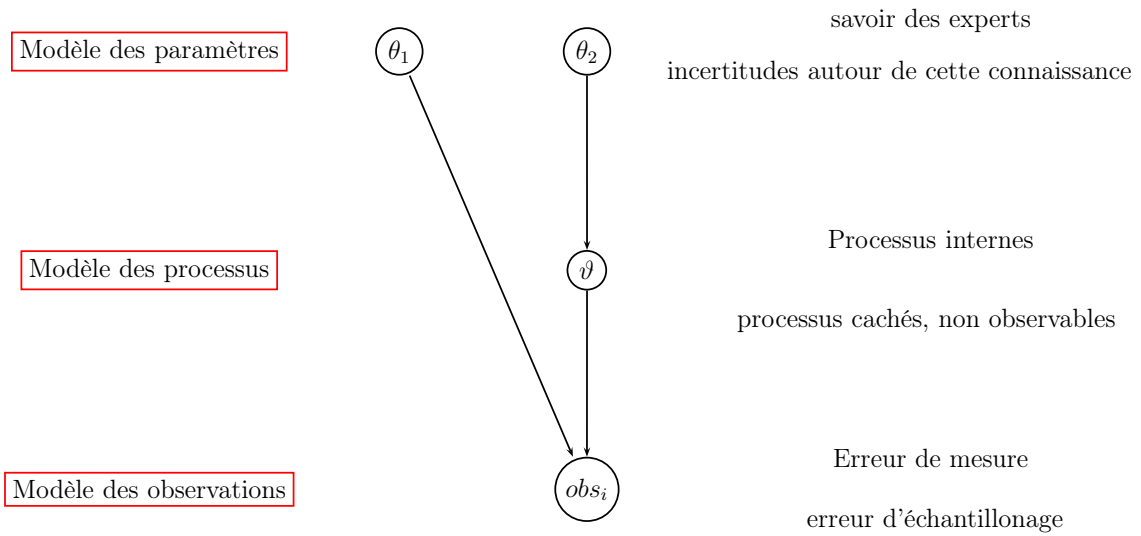
$$[obs|\vartheta, \theta_1]$$

2. Modèle des processus :

$$[\vartheta|\theta_2]$$

3. Modèle des paramètres :

$$[\theta_1, \theta_2]$$



# Le principe général et représentation graphique

1. Modèle des données :

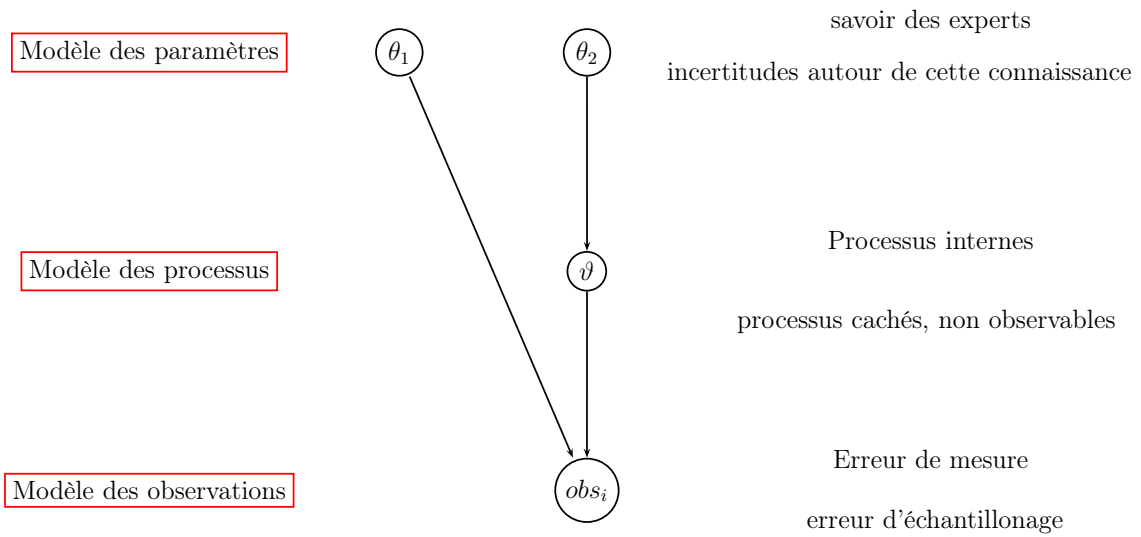
$$[obs|\vartheta, \theta_1]$$

2. Modèle des processus :

$$[\vartheta|\theta_2]$$

3. Modèle des paramètres :

$$[\theta_1, \theta_2]$$



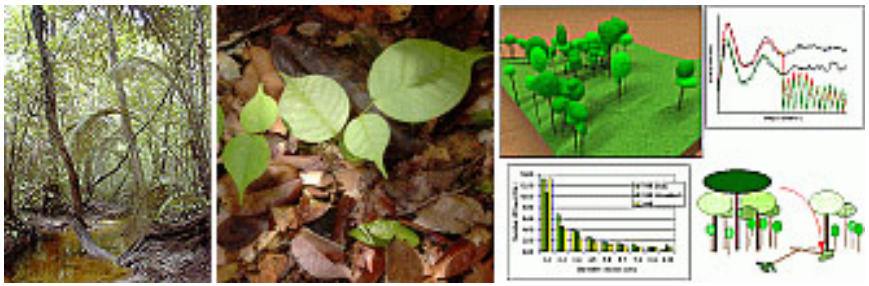

On s'intéresse in fine à la distribution a posteriori :

$$[\vartheta, \theta_2, \theta_1|obs] = \frac{[obs|\vartheta, \theta_1][\vartheta|\theta_2][\theta_1, \theta_2]}{\int [obs|\vartheta, \theta_1][\vartheta|\theta_2][\theta_1, \theta_2] d\vartheta d\theta_2 d\theta 1}$$

# L'écologie forestière

- Gestion durable des ressources forestières
  - Comprendre le fonctionnement des écosystèmes forestiers
  - Connaître l'évolution du peuplement
- Recours à la modélisation
- Modèles de dynamique forestière
  - Modèles individus centrés
  - Régénération, croissance, mortalité

➠ Talon d'Achille : prédiction de la régénération



# Modélisation de la régénération

1. Caractériser la dispersion des graines et du pollen
   - génotypes des juvéniles et des reproducteurs
   - géoréférencés
2. L'environnement influence la répartition spatiale
   - Altitude, teneur en azote,... (continue)
   - granulométrie (comptage)
   - intensité de la couleur (ordinale)

# Deux Questions

1. modélisation d'un champ spatial non-gaussien multivarié
2. prise en compte des erreurs de prédiction de l'environnement sur les paramètres de dispersion

# Modélisation de l'environnement [CMPB11]

$Y_1(\mathbf{s})$, la teneur en azote, variable gaussienne
$Y_2(\mathbf{s})$, la granulométrie, variable de Poisson

# Modélisation de l'environnement [CMPB11]

$Y_1(\mathbf{s})$, la teneur en azote, variable gaussienne
$Y_2(\mathbf{s})$, la granulométrie, variable de Poisson

Introduction de variables latentes continues

$$\forall i = 1, \ldots, n$$
$$Y_1(\mathbf{x}_i)|S_1(\mathbf{x}_i), \mu_1, \nu_1 \sim \mathcal{N}(\mu_1 + S_1(\mathbf{x}_i), \nu_1^2)$$
$$Y_2(\mathbf{x}_i)|S_2(\mathbf{x}_i), \mu_2 \sim \mathcal{P}(\exp(\mu_2 + S_2(\mathbf{x}_i)))$$

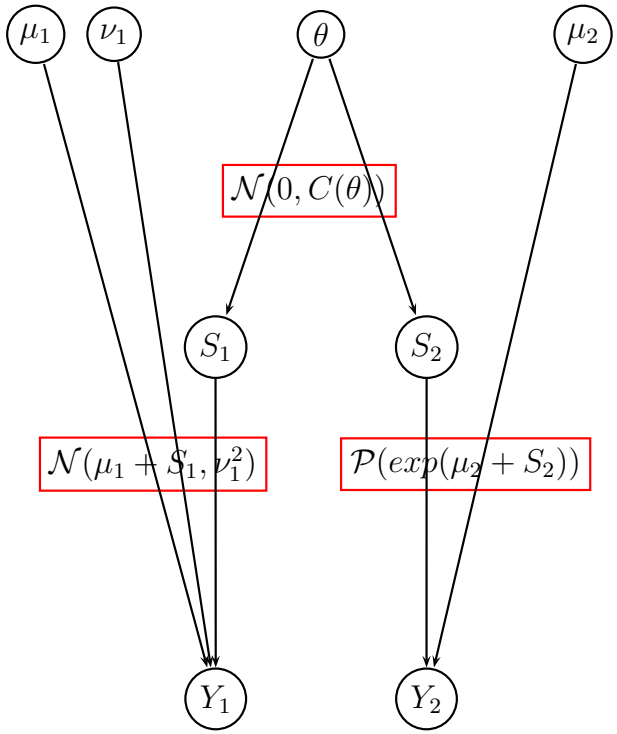
Modélisation de la dépendance spatiale entre les variables continues

$$\mathbf{S}_1, \mathbf{S}_2|\boldsymbol{\theta} \sim \mathcal{N}_{2n}(\mathbf{0}, \mathbf{C}(\theta))$$

Lois *a priori*

$$\boldsymbol{\theta}, \mu_1, \mu_2, \nu_1 \sim \text{ loi } a \textit{ priori}$$

# Représentation graphique



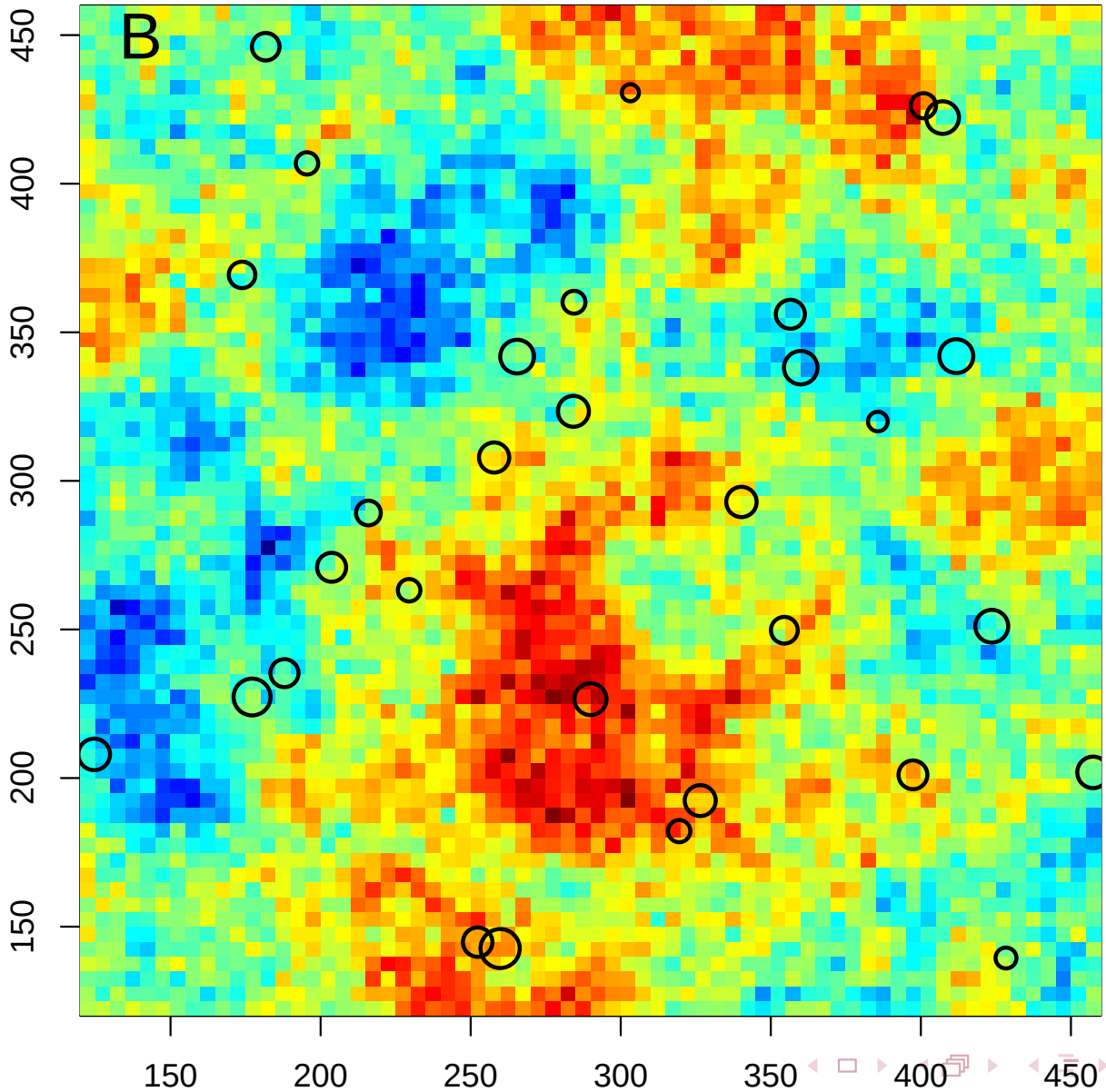
# Quelques remarques

- modèle spatial multivarié hiérarchique permet de prédire des variables de différente nature
- la prise en compte de la corrélation entre les variables grâce à une procédure d'estimation multivariée permet d'améliorer la qualité des prédictions

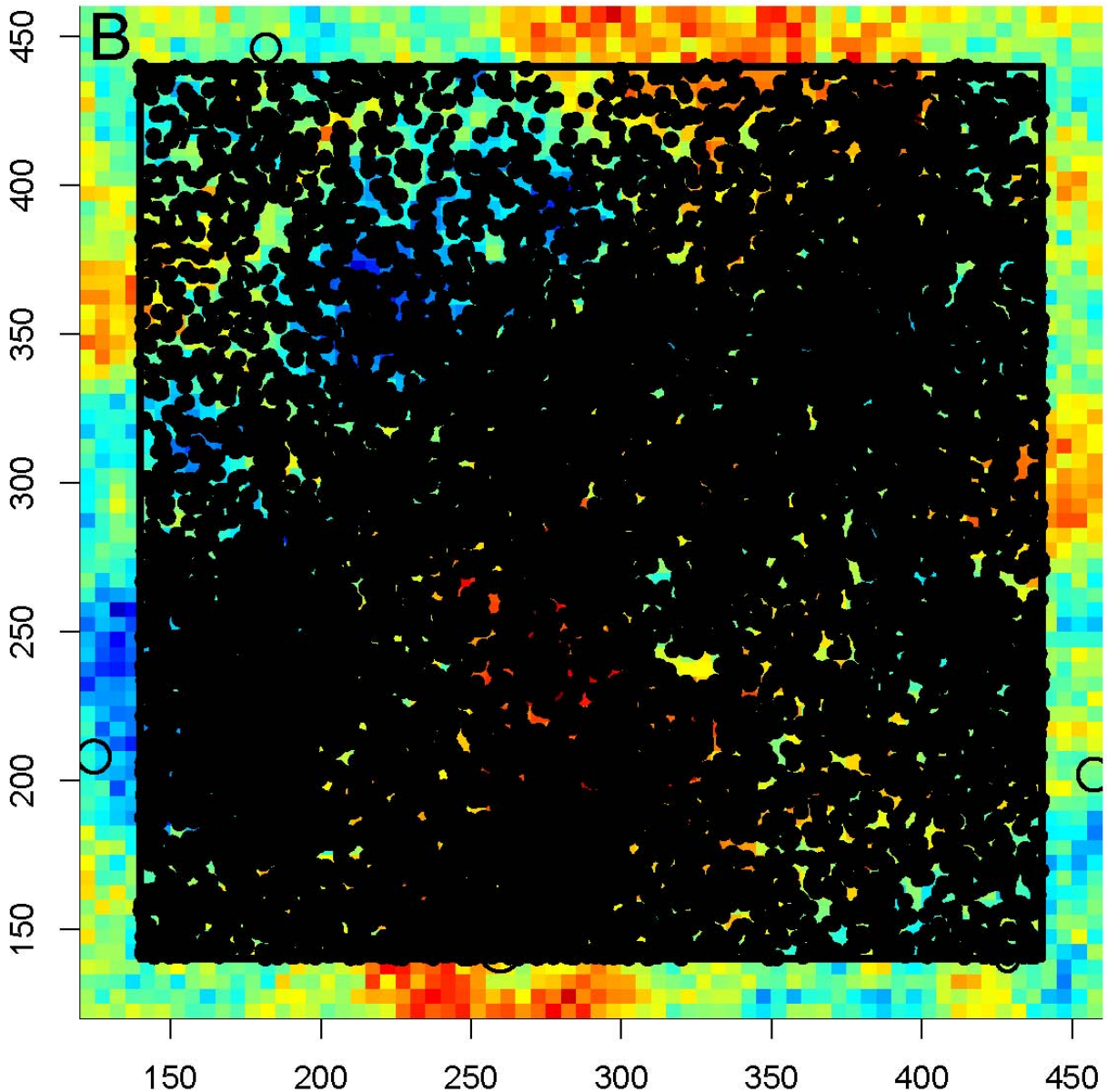
# Quelques remarques

- modèle spatial multivarié hiérarchique permet de prédire des variables de différente nature
- la prise en compte de la corrélation entre les variables grâce à une procédure d'estimation multivariée permet d'améliorer la qualité des prédictions
- procédure d'estimation des paramètres est gourmande en ressources informatiques
- la complexité augmente rapidement avec le nombre de variables

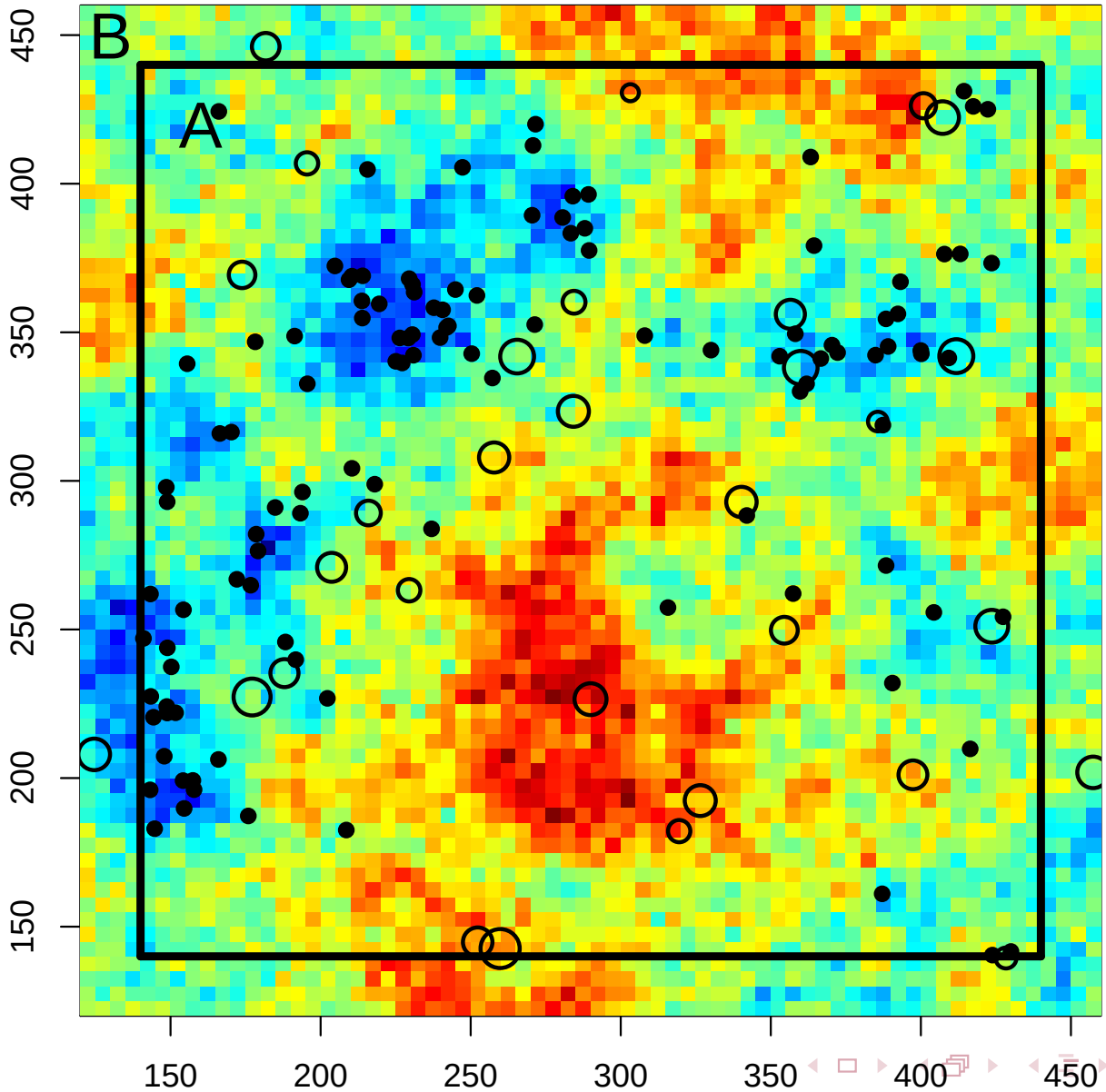
# Modélisation de la régénération



# Modélisation de la régénération



# Modélisation de la régénération



# Modèle de régénération

Soit $\mathbf{Y}(x)$ le vecteur des variables environnementales au point $x$
$(x_i, G_i), i = 1, \ldots, n$ est la réalisation d'un Processus de Poisson hétérogène marqué, d'intensité

$$\lambda_{\mathcal{S}}(x) = \sum_{G} \lambda_{\mathcal{S}}(x, G)$$

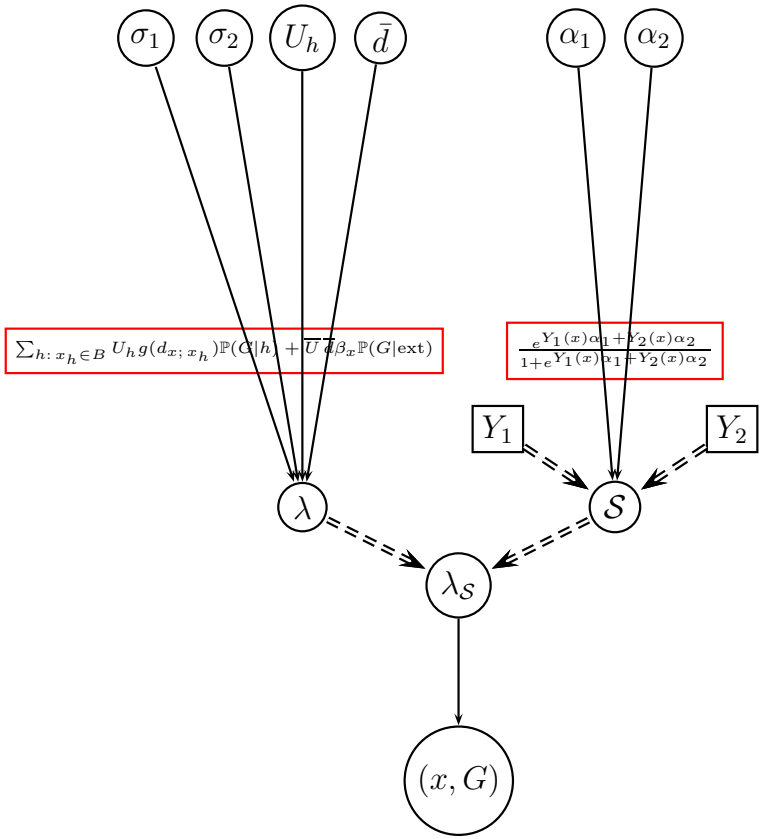
où

$$\lambda_{\mathcal{S}}(x, G) = \underbrace{\lambda(x, G)}_{\text{processus de dispersion}} \underbrace{\mathcal{S}(\mathbf{Y}(x))}_{\text{filtre environnemental}}$$

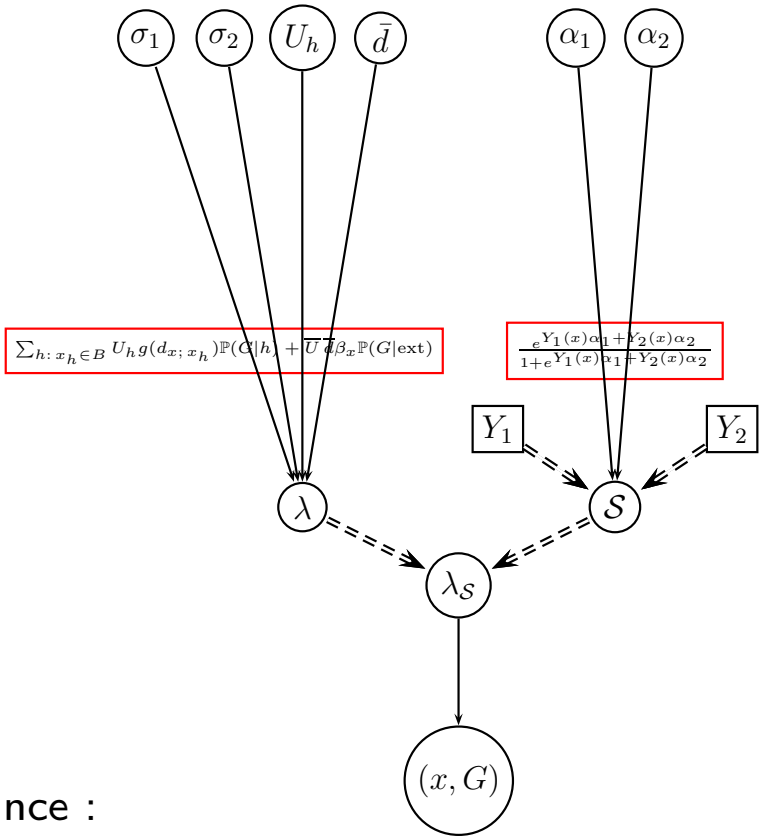
$$= \left( \sum_{h:\, x_h \in B} U_h g(d_{x;\, x_h}) \mathbb{P}(G|h) + \overline{U}\, \overline{d} \beta_x \mathbb{P}(G|\text{ext}) \right) \mathcal{S}(x)$$

où $\mathcal{S}(x) = \frac{\exp(\alpha_1 Y_1(x) + \alpha_2 Y_2(x))}{1 + \exp(\alpha_1 Y_1(x) + \alpha_2 Y_2(x))}$

# Représentation graphique



# Représentation graphique



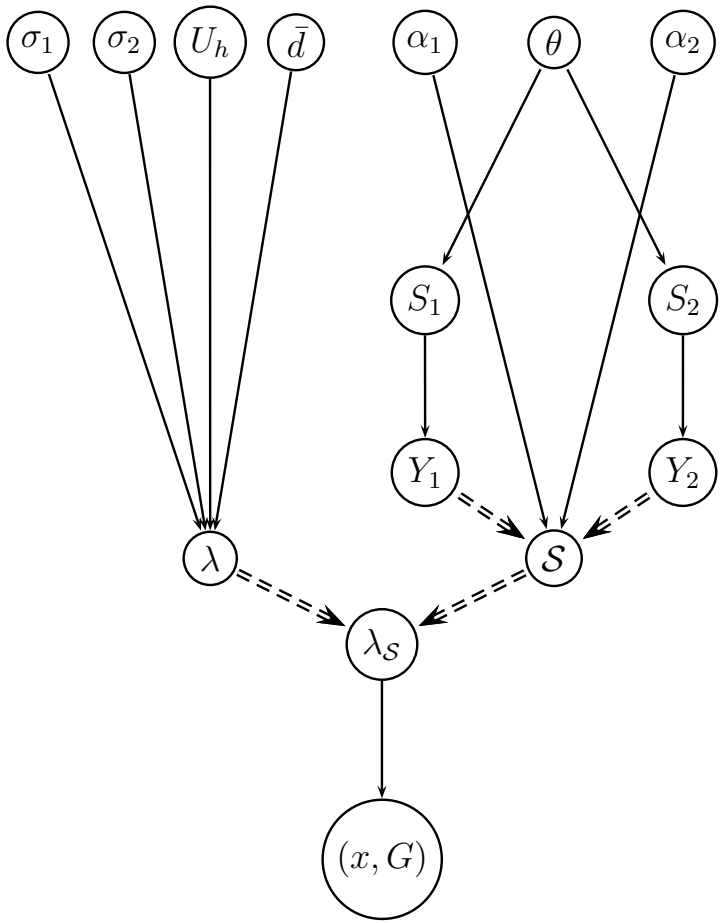

Estimation

- Maximum de vraisemblance :

$$\ell(\boldsymbol{\eta}; x_1, \ldots, x_n, \mathbf{Y}(x)) = \sum_{i=1}^{n} \ln(\lambda_{\mathcal{S}}(x_i, G_i; \boldsymbol{\eta})) - \int_A \lambda_{\mathcal{S}}(x; \boldsymbol{\eta}) \mathrm{d}x$$

- par méthode MCMC

# Prise en compte des erreurs de prédiction de l'environnement : représentation graphique



# Prise en compte des erreurs de prédiction

**Simulation**

Vraie valeur : $\sigma_1 = 95$

| Environnement | Estimation | Écart-type |
|---|---|---|
| Parfaitement connu | 95,76 | 4,12 |
| échantillonnage aléatoire + Préd. aux points où se trouvent les juvéniles | 101,49 | 6,32 |

- Augmentation du biais
- Augmentation de l'écart-type des estimations

# Remarques I

## Premiers résultats

- Estimation satisfaisante des distances de dispersion
  - si assez d'information génétique
  - si milieu non perturbé
- Problème d'estimation de la densité d'adultes reproducteurs

## Prise en compte des erreurs de prédiction

- Prise en compte directe des erreurs grâce à l'approche hiérarchique
- Augmentation de l'intervalle de crédibilité
- Nécessité de prendre en compte les erreurs de prédiction

# Remarques II

D'un point de vue Hierarchique

- Position et génotype des parents issus d'un processus ponctuel marqué
- Données génotypiques entachées d'erreur.

# Conclusions et perspectives I

1. Outil performant pour traiter des problèmes complexes
2. Permet de coupler simplement différents modèles
3. Cadre conceptuel homogène

## d'un point de vue Technique

- Inférence :
  - Dimension élevée, parallélisation
  - Outil de diagnostique de convergence
- Comparaison de modèles, Facteur de Bayes, Deviance Information Criterion (DIC) [SBCvdL02],
- Adéquation aux données, posterior predictive check (ppc) [GMS96]

# Conclusions et perspectives II

## Nécessité de travailler en collaboration

- Formalisation des processus cachés par les scientifiques du domaine
- Développer un même language (approche graphique peut aider)
- Les échanges entre chercheurs et statisticiens doivent être permanents

# Références I


N. Cressie, C.A. Caldery, J.S. Clark, J.M. Ver Hoef, and C.K. Wikle.

Accounting for uncertainty in ecological analysis : the strengths and limitations of hierarchical statistical modeling.
*Agencies and staff of the US department of commerce*, 2009.

J.S. Clark and A.E. Gelfand.

A future for models and data in environmental science.
*Trends in Ecology & Evolution*, 21(7) :375–380, 2006.

J. Clark.

Why environmental scientists are becoming bayesians.
*Ecology Letters*, 8 :2, 2005.

P. Chagneau, F. Mortier, N. Picard, and J.N. Bacro.

Hierarchical bayesian model for spatial prediction of multivariate non-gaussian random fields.
*Biometrics*, 67 :97, 2011.

A. M. Ellison.

Bayesian inference in ecology.
*Ecology Letters*, 7 :509–520, 2004.

Andrew Gelman, Xiao-Li Meng, and Hal Stern.

Posterior predictive assessment of model fitness via realized discrepancies. (With discussion).
*Stat. Sin.*, 6 :733–807, 1996.

S.M. McMahon and J.M. Diez.

Scales of association : hierarchical linear models and the measurement of ecological systems.
*Ecology Letters*, 10(6) :437–452, 2007.

# Références II


E. Parent and J. Bernier.

***Le raisonnement bayésien. Modélisation et inférence.***
Springer, 2007.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde.

**Bayesian measures of model complexity and fit (with discussion).**
*J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 64 :583–639, 2002.

Christopher K. Wikle.

**Hierarchical Models in Environmental Science.**
*International Statistical Review*, 71 :181–199, 2003.

C.K. Wikle and J. A. Royle.

**Spatial Statistical Modeling in Biology.**
*Encyclopedia of Life Support Systems, EOLSS Publishers Co. Ltd.*, 2004.